SEEK, AND YE SHALL FIND

Robert Filman • *Lockheed Martin* • *filman@mcc.com*
Feniosky Peña-Mora • *MIT* • *feniosky@mit.edu*

In the spirit of this issue, the Spider went on a tour of search engines.

You might think that search engines are just for search, but engine purveyors want to be your home page on the Web. Their pages tend to be decorated with directories, links to atlases, yellow pages, news, "favorite sites," and, of course, advertisements.

How can we compare search engines? Only arbitrarily. In an unscientific sampling technique, the Spider tried the search engines with the following goals:

- *Magna Carta*. Find the text of the Magna Carta. This should be an easy task. Points here for getting to a page that has the text and for ranking the page high among the links found.
- *Mercury*. The specific gravity of Mercury. Here we're looking for a factoid likely to be buried in something else.
- *Hide and Seek*. Here we're looking for a specific image, a painting by surrealist Pavel Tchelitchew called "Hide and Seek." (Bob wanted to use this as the cover for the issue, but minor considerations like copyright laws and economics got in the way. You can check it out by following the link on **http://www.olemiss.edu/~slarson/isit/oldstuff/oldresults.html**.) Effectively, this is the only right answer we found on the Net. (The real picture, which hangs in the Museum of Modern Art in New York, is far more impressive than the jpeg.)
- *Publish*. This is an amorphous search request for the software topic "event-based programming using publish and subscribe." We're looking for technical information, not marketing hype. This is a particularly difficult search, because the search terms all have meanings in many other domains.
- *Feniosky*. This is a people-finding task. Since there's only one Feniosky, finding him should be straightforward.
- *McCarthy*. This is the opposite people-search problem. "John McCarthy" is a common name. We check which additional clues are needed to guide the search engine to Stanford and John's work.

The results for each engine are listed in the sidebars on the following pages.

Flies have been awarded by an absolutely precise but utterly arbitrary system. Since the Tourist found no real differences in quality of user interface and didn't perform a consistent enough set of experiments to determine speed, fly parts are achieved by

- weighting each question,
- giving each search engine a score for question, and
- computing the weighted sum.

For those who want to quibble, the specific values used are listed in Table 1. For space considerations, most of the pointers to the sites found in our searches have been banished to the online version of the column (**http://www.computer.org/internet/v2n4/w4arach.htm**). Think of this omission as a good reason to check out the online magazine.

Lycos · www.lycos.com

Lycos presents a crowded home page. Among its features are 22 Web guides, each with three topics. In contrast to Dewey, the topics and subtopics seem not to be a comprehensive distribution of human knowledge, but rather based on, well, what someone at Lycos must have thought people would be looking for. For example, the Web guides include the topic "Space/Sci-Fi" with the subtopics "Exploration," "X-Files," and "Star Trek," but no topic for "Science." There are also pointers to specialized search engines for things like books, road maps, city guides, and yellow pages; headlines for three news stories; and various shopping opportunities.

The primary search function offers a dozen choices of places to search (the Web, Web reviews, home pages, news, sounds, pictures, and so on), help, and advanced search. Advanced search includes disjunction, "natural language query," conjunction (including ordering and proximity relationships), and exact phrases. The user can also influence sorting by choosing where in the text the words appear.

Overall Lycos did well at finding people, but seemed weaker at both breadth of coverage and precision of recalled topics. The Lycos approach of looking for proximate words without being able to exclude topics found a lot of stuff. Occasionally, some of it was relevant. (*3.17 Flies*)

infoseek · www.infoseek.com

Infoseek's top level splits the page vertically—one side for search and the other for 17 "channels" (each with 2–4 subtopics). Once again, the channels must have been based on marketing experience rather than a comprehensive sort of human knowledge—we see "Entertainment" followed by "The Good Life," though no channel for "History." Infoseek's search vocabulary allows one to demand terms (+Magna) while discouraging others (–rock). The pipe operator "|" can be used to progressively narrow the search; other operators can be used to search backwards over links, and with respect to sites, titles, and URLs. Capitalization is used to infer multiword proper nouns.

Infoseek groups results by site, allowing one to request "more results from that site." This probably works better when the site is a small organization and not AOL. The advanced search mode is a dialog-box–based interface for the more expressive search language.

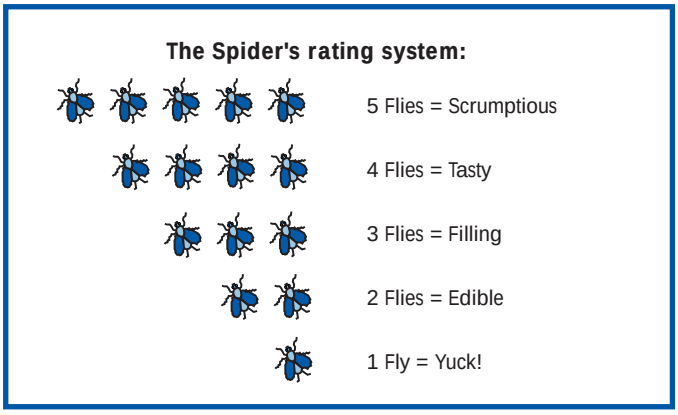
We liked the ability to successively focus queries with |, though we were disappointed that infoseek didn't do better with the more difficult textual query. (*3.83 Flies*)

Excite · www.excite.com

Excite's twist on the world seems to be to customize to particular users. After having registered, the Spider was able to adjust the top-level page to details of interest (such as choosing columnists and cartoons) while excluding others (like horoscope).

Turning to search, Excite has a dozen nodes in its top-level directory structure. We clicked from "Computers and Internet" through "Computer Science" on to "Distributed Computing" and then did a search on "publish and subscribe," which yielded nothing in the first 20 results. For free-form searches, Excite claims to do not merely word matching, but concept-based search, "looking for ideas closely related to the words in your query." The advanced search language includes double-quoted phrases, plus and minus prefixes for "must" and "cannot," Boolean operators (including parentheses), and a search wizard that suggests terms to add to the query. The "Power Search" link offers the ability to search the Web or news, a bit of fill-in-the-blanks for those who can't write Boolean expressions, and a touch of control over the results presentation.

Overall, we'd call Excite's query language minimalistically adequate. That is, it had all the operations we'd demand in a query language (particularly the ability to require and exclude terms) but not a whole lot more. Interestingly, Excite did well on the nebulous research question, but stumbled on the simpler, more directed queries. (*3.74 Flies*)



WebCrawler · www.webcrawler.com

Corporate Excite includes Excite's original search engine, WebCrawler, and the Magellan directory search system. WebCrawler presents a dozen and a half channels in its directory structure and the opportunity to personalize the presentation. When we went to personalize, we discovered our Excite profile already resident.

WebCrawler has a straightforward query language, with double quotes for specific phrases, Boolean operators, parentheses, + and –. Overall, A minimally adequate query language, an excellent job on the ill-specified research query, but relatively weak on specific facts. (*3.72 Flies*)

HotBot · www.hotbot.com/

HotBot is more a search engine and less a place to live. The home page offers 12 directories, a dozen specific search places (for example, yellow pages, stocks, shareware), collections of *Wired* articles and online shopping, but noticeably, no attractions (or distractions) for the nonsearcher. The

Table 1. The specific values used for scoring the search engines tested and the resulting scores.

Task	Weight	Lycos	infoseek	Excite	WebCrawler	HotBot	AltaVista
Magna	0.17	2	5	5	5	0	4
Mercury	0.17	4	4	2	0	5	3
Tchelitchew	0.17	0	5	1	0	5	4
Publish	0.40	3	2	5	5	1	4
Feniosky	0.08	5	4	2	4	5	2
McCarthy	0.11	5	3	2	5	5	2
Total Score		**3.17**	**3.83**	**3.74**	**3.72**	**3.05**	**3.85**

LOOKING FOR A SPECIFIC TEXT: THE MAGNA CARTA

Lycos	A simple request for 'Magna Carta' in the plain search engine window didn't list the document text until the 19th item, and that, at the British Library, turned out to be a zoomable view of the original document. In an advanced search, we asked for '"Magna Carta" text,' in any order, in English, in the entire document (we would have limited it to the title, but feared "text" would throw it off), and made a few arbitrary guesses on the textual arrangement of the key. This gave us the text as the sixth item and again as the ninth. Lycos doesn't have any negation operator, so we couldn't exclude the obviously nongermane uses of Magna Carta (the rock band).
infoseek	The first infoseek reference went straight to the Federal National Archives and Records Administration, which includes not only the text but also an essay on its importance in history. The eighth reference was to the British Library. A punch parody on the deprecation of the Magna Carta in British law also made the list. The remaining references were to the rock band "Magna Carta." Clarifying the request to 'Magna Carta –rock –band' thinned the musical references but didn't eliminate all of them.
Excite	Our request for '"Magna Carta"' in the plain search engine window turned up the NARA site as the first reference, and mixed rock band and political references about equally in the rest. "More links like this one" from the NARA site found other relevant pages, such as the Mississippi State archive to US history documents.
WebCrawler	A search for '"Magna Carta"' yielded the NARA site as the first reference, rock bands for the next five, and then four more political documents. '"Magna Carta' –rock –band' raised the density of relevant hits to 70 percent.
HotBot	A simple search for 'Magna Carta' revealed much about the band and little about the document. An advanced search for 'Magna Carta' but not 'rock' or 'band' didn't seem to change things. Evidently, in HotBot, the operator "must not contain" is only a suggestion.
AltaVista	Entering 'Magna Carta' in the search engine window got the original document as the second and seventh items. Other links returned by AltaVista pointed to a variety of Web sites connected with Magna Carta—from the Magna Carta society in New Zealand to a Magna Carta book at Amazon.com. AltaVista was pleased to inform us that "about 176023 matches were found," so we couldn't check them all out.

top-level search box offers searching with respect to the key ("all the words," "any of the words," "title," "Boolean phrase," and so on), temporal and domain limits (such as, "look only in .gov"), media requirements (for example, image) and number and style for results. Advanced search is a dialog box interface to the above, plus the ability to specify a wider variety of search objects (such as VRML and ActiveX). We had to add a "?" to the end of the search path to reach HotBot from a Mac.

HotBot seems to have a broader reach than the previous engines. For well-focused questions, it found good answers. For queries where lots of things could match, well, the precision and relevance were a little weak. (*3.05 Flies*)

AltaVista · altavista.digital.com/

AltaVista is a service provided by Compaq. It originated with Digital Equipment Corp. as a showcase for the computational ability of the DEC Alpha computer but has evolved into a search-engine business of its own. The AltaVista home page combines the search function with both specific services for finding people and companies and "Zones," which seem to be interesting articles highlighted by the service. One novel feature is an automatic translation service by Systran. The translation service is connected to the search facility to allow single-click translation of found pages. Continuing this international spirit, AltaVista can also search in Chinese, Japanese, and Korean.

AltaVista has a very large database with references to a great variety of online documents ranging from Web pages to mailing lists. This may result in very broad and quasi-exhaustive results that need to be narrowed with its more advanced search features.

Thus, if it is out there, AltaVista will find it. However, unless you use the advanced search features, the results tend to be overwhelming. (*3.85 Flies*)

Common sense and lack of space keep us from visiting the other search engines out there (like LookSmart, Snap!, Northern Light, Whatuseek, and GoTo). Check them out yourself. Try "WWW Search Engines" in your favorite search engine and see if they can find the competition.

What can we conclude about all this? Read the directions of the search engine to make sure you've got the syntax right. Try playing around with different command strings—when ranking 10 million matching pages, a small shift in the query will bring a different set of pages to the top. For vague, researchy queries, try a number of engines—the output is often radically different.

SEARCHING FOR FACTOIDS: THE SPECIFIC GRAVITY OF MERCURY

Lycos We asked the top-level engine 'the specific gravity of mercury.' In the eighth entry, a US government occupational health guideline, we found an answer: 13.5. An advanced search that sought high proximity at the cost of early mention pretty much found the same documents but also gave a link to the Lycos "reference room," where a couple of straightforward clicks led to 13.546.

infoseek We tried '"specific gravity" mercury.' The third reference (just a short paragraph) revealed the desired factoid, the answer itself showing on the search page. None of the other top 20 items seemed to have the information, though we were certainly intrigued to find a page on "Simple Way to Beat Urine Tests" for drugs.

Excite We entered '"specific gravity" mercury' into the top-level engine. The third entry was a Purdue physics lab page that mentioned the value (13.6), but it was the only one in the top 20. Taking the suggestion to add 'mass' to the search produced a lot more physics, but not much of it particularly focused on the question.

WebCrawler The query '"specific gravity" mercury' produced a lot of data on the physics on the planet Mercury (including a handy reference to a site listing Elvis's weight on other planets), but zilch on density. Trying to focus a little more specifically, we demanded '+"specific gravity" +mercury.' This dramatically changed the results, with more pages about gold mining, fish culture, time-travel machines, and heavy-metal bands, but nothing in the 13.5 range.

HotBot Demanding both "specific gravity" and "mercury" found on the second reference a collection of conversion factors including a specific gravity of 13.5951 at 0 degrees C, and on the third, a table of specific gravities, yielding 13.546.

AltaVista We tried '"specific gravity of Mercury".' The first link pointed to a Web site at the University of Michigan that had the desired fact buried in its text; the second and third links also contained the desired values. Trying '"specific gravity" mercury' finds only one reference in the top 10.

IMAGE SEARCH: TCHELITCHEW'S "HIDE AND SEEK"

Lycos A top-level search for 'Tchelitchew "Hide and Seek"' got no results. A picture-directed advanced search, with emphasis on proximity, didn't do any better.

infoseek 'Tchelitchew, "Hide and Seek"' produced an abundance of references, the first being our target picture and the rest ranging all over the place. Rephrasing the query as 'Tchelitchew | "Hide and Seek"' produced a single bull's eye. Note that if we hadn't read the infoseek directions, we might have asked for 'Tchelitchew Hide and Seek' which finds almost nothing on the person "Tchelitchew Hide" and a whole lot on seeking, including most competitive search engines.

Excite A top-level search for 'Tchelitchew "Hide and Seek"' found many references to hide and seek but none to Tchelitchew. Trying '+Tchelitchew "Hide and Seek"' got us the Electric Library's Encyclopedia, and several other pages that mentioned Tchelitchew, but no images of his work.

WebCrawler 'Tchelitchew "Hide and Seek"' was long on hide and seek but short on Tchelitchew. Trying '+Tchelitchew "Hide and Seek"' produced two references that mentioned Tchelitchew in passing.

HotBot A simple search for 'Tchelitchew "Hide and Seek"' found four references, the first of which was the target. An advanced search that required an image and both 'Tchelitchew' and the phrase "Hide and Seek" found three links. Once again, the first was the target. Seeking other images by Tchelitchew found 59 links, one of which is the paper "Can Surgery Repair an Osteoporotic Spine?" which includes a Tchelitchew painting much in the style of "Hide and Seek." That search also revealed photographs of Tchelitchew in the Library of Congress.

AltaVista We decided to use AltaVista's advanced search features for this query, entering "Tchelitchew" AND "Hide and Seek." Instead of searching for an image name, we searched the entire text. The search gave two results, the second of which was the target.

TOPIC SEARCH: ZEROING IN ON THE MEANING YOU SEEK

Lycos When we typed 'publish and subscribe event based programming papers' in the top-level engine, we got one match, which 404'd. An advanced search, which favored frequency over matching all the terms, found 72 links. These included: item 5, a list of CORBA products, including one with a publish and subscribe service; item 6, a chapter by Grady Booch on his eponymous method; and item 17, a list of Java products including Open Horizon's Ambrosia. Just '"publish and subscribe"' with weight toward the beginning of the document found a World Wide Web Consortium workshop on push technology, the object-orientation FAQ, an article urging the use of software agents to reduce Net traffic, along with some thoroughly off-the-wall pages, such as Net classes on religion and a court ruling on compensation for vaccine-induced injury. Putting 'publish and subscribe' in double quotes got a different set of links with a slightly lower density of interesting pages.

infoseek We tried 'programming | event publish subscribe.' While there were many interesting references in the list, only the Sun Java site appeared in the top 20. We tried 'programming | languages | event-based "publish and subscribe,"' which got us four links, the fourth of which was a glossary item from BEA Systems defining publish and subscribe. 'Computers | "publish and subscribe"' yielded little of value, and 'programming | "publish and subscribe"' added a case study on the use of publish and subscribe and a pointer to a Neuron Data tool.

Excite 'Computers programming languages event +"publish and subscribe"' gave us: item 1, a TIBCO press release; 2, a description of the Tuxedo System EventBroker; 3, a Level 8 Systems press release on EventWorks; 6, a product description of Neosoft's MQIntegrator; 7, the Booch article; and 10, Tom Laffey of Talarian Corporation's paper "Publish-Subscribe Middleware for Real-Time, Event-Driven Computing." Pursuing the "more documents like this one" link on the last entry led to Sally Cusack's article "Rethinking Data Delivery" in the e-zine *Application Development Trends*. (This density of good reference probably earned a fly and two wings just by itself.)

WebCrawler We tried '+"publish and subscribe" computers event programming language systems.' In the first dozen results, we got: item 3, Neon; 6, the Distributed Objects mailing list FAQ, which included in its "people on the list" link a list of a dozen links to publish-and-subscribe technologies (bingo!); 7, a page from Berkeley on commercial message-oriented middleware; 8, BEA Tuxedo; 9, an Apple description of MacOS X, including publish-and-subscribe facilities; 11, Level 8 Systems; 12, ObjectSpace Voyager, a freeware Java ORB that includes publish-and-subscribe services. Even most of the misses were in the right neighborhood, (including the lovely Acronym Emporium—a list of 1,365 computer acronyms "for all your acronym needs"—a great resource to keep you from being left out during the next bull session at Fry's.

HotBot '"Publish and subscribe" programming language system computer event' as a simple search got a lot of stuff fairly far off the topic. Moving on to advanced search, we required "publish and subscribe" and recommended the other terms. This got us a different set of links, including a paper by Hall, Bates, and Bacon on "Flexible Distributed Programming Using Mobile Code," but nothing else of interest among the top 20 listings.

AltaVista The entry 'event-based programming using publish and subscribe' in the simple search mode returned good matches for items 1–5, 8, and 9, among the top 10 items. Trying 'event-based programming "publish and subscribe"' found four interesting links among the top 20. We tried the advanced feature of AltaVista using "event-based programming" AND publish AND service. This resulted in only one matching search result pointing to a newsletter devoted to geographic information systems.

METASEARCH

Well, if different search engines can find different results, why aren't there programs that try running several search engines in parallel, presenting the "best" answers? There are—the metasearch engines. The Spider crawled on over to one to see what it was like. (See also the Lawrence and Giles and Benitez et al. articles in this issue for more information about metasearch engines.)

Net7 from Datatrak · datatrak.net7.co.uk/

This site provides a metasearch applet that allows metasearches to be made over the Web, using Excite, Yahoo, Infoseek, AltaVista, Lycos, HotBot, WebCrawler, Magellan, Northern Light, Open Text; the Usenet news using DejaNews, AltaVista News, HotBot News; and headline news using HotBot NewsBot, Yahoo News, Excite News, Infoseek News.

Each search engine gets a separate window for its top results; Net7 neither imposes nor presents any comparison or ranking of the results on its own. It functions purely as a broadcast

PEOPLE SEARCH: FENIOSKY

Lycos	Typing 'Feniosky' into the simple search request went right to Feniosky's project, with the next dozen or so links being references to Feniosky's papers.
infoseek	The first listing on the results page was for one of Feniosky's students, with a "grouped results from" MIT that led to his home page. Most of the other results seemed to be indirect pointers to the Tourist.
Excite	A simple request for 'Feniosky' got references to papers and co-authors, but even the "more documents like this one" link didn't find his home page.
WebCrawler	'Feniosky' found two links, the first to Feniosky's group at MIT, and the second to an old conference announcement.
HotBot	A simple search for 'Feniosky' went right to Feniosky's home page. The next nine references (out of 85 total) were either to other MIT locations or *IEEE Internet Computing*.
AltaVista	We entered "Feniosky" into the simple search window and got 247 matches. AltaVista didn't show Feniosky's home page until the 20th link. There were hosts of other links pertaining to the Distributed Software Engineering Laboratory Mailing List Archive.

FINDING ONE AMONG MANY: JOHN MCCARTHY

Lycos	'John McCarthy' as a simple request found his faculty listing as the first and second items, and a McCarthy book as the fourth, with a link labeled "more like this" full of artificial intelligence references. What we can't figure out is how the machine knew we were looking for *that* John McCarthy.
infoseek	The fourth item was McCarthy's home page.
Excite	The sixth item on this list pointed to the Computer Science Department bio, which had a link to McCarthy's home page.
WebCrawler	Items 1, 2, and 6 pointed to McCarthy's home page and publications, and item 4 to an interactive Lisp tutorial. These surrounded a pointer to the Edgar Bergen and Charlie McCarthy Show and a reference to the Labyrinth Latin Library, which was puzzling as it registered as already visited. It turned out to have not only a rendition of the Magna Carta but also a translation of the Latin Vulgate Bible by Dennis McCarthy, illustrating that it's a smaller Web than we thought.
HotBot	The first reference was to McCarthy, the second to his group, and the remaining references split about evenly between bibliographic pointers and irrelevant links.
AltaVista	The results from the simple request "John McCarthy" ranged from John McCarthy in East Lansing, Michigan, to John McCarthy a car dealer. The first reference to John McCarthy from Stanford was a quotation in the seventh search result. We decided to nail down the search results by using Stanford as an additional parameter in the advanced search option. This option was more accurate, with the top five search results pointing to the right John McCarthy. We can conclude that AltaVista treats all John McCarthys equally.

interface to the underlying systems.

Lest this seem too much of a virtue, understand that Net7 is using a shotgun. Site-specific operators (like "narrow" or "search in the titles" or "search in English") can provide a more precise targeting. Thus, a metaengine like Net7 may prove more useful for finding the page that is specific and well defined but not necessarily known to every engine's crawler than it is to a broad topic-research query. That is, Net7 would have been helpful finding "Hide and Seek" but would have gotten in the way on "Publish and Subscribe."

About the Spider

The Arachnoid Tourist scours the Net to find and review Web sites of interest to our readers.

Contact Robert Filman at filman@atc.lmco.com, or Feniosky Peña-Mora at feniosky@mit.edu, or the magazine at internet-computing@computer.org.